Incremental Spectral Clustering with the Normalised Laplacian

Summary

Position

- Spectral clustering approach ...
  - Eigen-decomposition of a (Laplacian) matrix
  - k-means on eigen-vectors
- ... on (slowly) changing graphs

$\Rightarrow$ Exact computation at each iteration is too expensive

Proposed approach

- Fast update of the eigen-decomposition / of the clustering

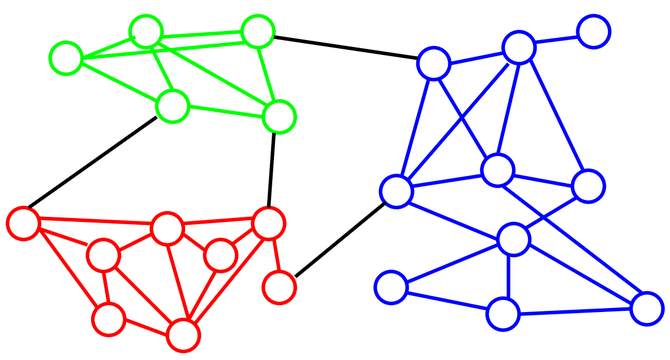
Charanpal Dhanjal[1], Romaric Gaudel[1,2], Stéphan Clémençon[1]

[1] LTCI (UMR 5141, Télécom ParisTech/CNRS), Paris, F-59653

[2] LIFL (UMR 8022, univ. Lille/CNRS) & INRIA Lille Nord-Europe, Lille, F-59653

Standard spectral clustering [2] ...

Aim: group similar nodes of a graph



Spectral approach: eigen-decomposition + k-means

- $\mathbf{W}$: adjacency matrix (symmetric)
- $\mathbf{D}$: degree matrix (diagonal, $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$)
- $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$: Normalized Laplacian matrix
- $\mathbf{Q}\mathbf{\Omega}\mathbf{Q}^T = \mathbf{L}$: eigen-decomposition of $\mathbf{L}$
- $\mathbf{Q}_k$: eigen-vectors (columns in $\mathbf{Q}$) corresponding to the k smallest eigen-values
- $\tilde{\mathbf{Q}}_k$: $\mathbf{Q}_k$ normalized row by row
- $\mathbf{c}$: cluster membership vector, obtained through k-means($\tilde{\mathbf{Q}}_k$)

... in the context of changing graphs

- $\mathbf{W}$ depends on time t ($\mathbf{W}^{(t)}$)
- $\mathbf{W}^{(t)} \approx \mathbf{W}^{(t+1)}$
- $\mathbf{W}^{(t)}$ big (thousands of nodes)

$\Rightarrow$ Exact computation of $\mathbf{Q}_k^{(t)}$ and $\mathbf{c}^{(t)}$ at each iteration is too expensive

$\Longrightarrow$ Update $\mathbf{Q}_k^{(t)}$ and $\mathbf{c}^{(t)}$
leveraging previous eigen-decompositions and clusterings

Update of the best rank-k approximation

(looking for the greatest eigen-values instead of the smallest one)

Aim

- Known: $\mathbf{Q}_k^{(t)}\mathbf{\Omega}_k^{(t)}\mathbf{Q}_k^{(t)^T}$ best rank-k approximation of $\mathbf{L}^{(t)}$
- To compute: $\mathbf{Q}_k^{(t+1)}\mathbf{\Omega}_k^{(t+1)}\mathbf{Q}_k^{(t+1)^T}$ best rank-k approximation of $\mathbf{L}^{(t+1)}$
  - $\mathbf{L}^{(t+1)} = \mathbf{L}^{(t)} + \mathbf{U}^{(t+1)}$
  - $\mathbf{U}^{(t+1)} = \mathbf{Y}_1\mathbf{Y}_2^T + \mathbf{Y}_2\mathbf{Y}_1^T$ (small rank)
  - $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^{n\times p}$
- Proposed approach: return the best rank-k approximation of $\mathbf{Q}_k^{(t)}\mathbf{\Omega}_k^{(t)}\mathbf{Q}_k^{(t)^T} + \mathbf{U}^{(t+1)}$

Living space of $\mathbf{Y}_1$

- Deflate $\mathbf{Y}_1$
  - $\bar{\mathbf{Y}}_1 = (\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^T)\mathbf{Y}_1$
- Singular-decomposition of $\bar{\mathbf{Y}}_1$
  - $\bar{\mathbf{P}}_1\bar{\mathbf{\Sigma}}_1\bar{\mathbf{Q}}_1^T$

$\Rightarrow \mathbf{Y}_1$ lives in $\mathrm{Span}(\mathbf{Q}_k^{(t)},\ \bar{\mathbf{P}}_1)$

Living space of $\mathbf{Y}_2$

- Deflate $\mathbf{Y}_2$
  - $\bar{\mathbf{Y}}_2 = (\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^T - \bar{\mathbf{P}}_1\bar{\mathbf{P}}_1^T)\mathbf{Y}_2$
- Singular-decomposition of $\bar{\mathbf{Y}}_2$
  - $\bar{\mathbf{P}}_2\bar{\mathbf{\Sigma}}_2\bar{\mathbf{Q}}_2^T$

$\Rightarrow \mathbf{Y}_2$ lives in $\mathrm{Span}(\mathbf{Q}_k^{(t)},\ \bar{\mathbf{P}}_1,\ \bar{\mathbf{P}}_2)$

Decomposition of $\mathbf{Q}_k^{(t)}\mathbf{\Omega}_k^{(t)}\mathbf{Q}_k^{(t)^T} + \mathbf{U}^{(t+1)}$

- $\mathbf{Q}_k^{(t)}\mathbf{\Omega}_k^{(t)}\mathbf{Q}_k^{(t)^T} + \mathbf{U}^{(t+1)} = \tilde{\mathbf{Q}}\mathbf{\Delta}\tilde{\mathbf{Q}}^T$
  - $\tilde{\mathbf{Q}} = [\mathbf{Q}_k\ \bar{\mathbf{P}}_1\ \bar{\mathbf{P}}_2] \in \mathbb{R}^{n\times(k+2p)}$
  - $\mathbf{\Delta} = \tilde{\mathbf{Q}}^T(\mathbf{Q}_k^{(t)}\mathbf{\Omega}_k^{(t)}\mathbf{Q}_k^{(t)^T} + \mathbf{U}^{(t+1)})\tilde{\mathbf{Q}}$

$$= \begin{bmatrix} \mathbf{\Omega}_k + \mathbf{Q}_k^T\mathbf{U}\mathbf{Q}_k & \mathbf{Q}_k^T\mathbf{U}\bar{\mathbf{P}}_1 & \mathbf{Q}_k^T\mathbf{Y}_1\bar{\mathbf{Q}}_2\bar{\mathbf{\Sigma}}_2 \\ \bar{\mathbf{P}}_1^T\mathbf{U}\mathbf{Q}_k & \bar{\mathbf{P}}_1^T\mathbf{U}\bar{\mathbf{P}}_1 & \bar{\mathbf{\Sigma}}_1\bar{\mathbf{Q}}_1^T\bar{\mathbf{Q}}_2\bar{\mathbf{\Sigma}}_2 \\ \bar{\mathbf{\Sigma}}_2\bar{\mathbf{Q}}_2^T\mathbf{Y}_1^T\mathbf{Q}_k & \bar{\mathbf{\Sigma}}_2\bar{\mathbf{Q}}_2^T\bar{\mathbf{Q}}_1\bar{\mathbf{\Sigma}}_1 & \mathbf{0} \end{bmatrix}$$

- $\mathbf{H}_k\mathbf{\Pi}_k\mathbf{H}_k^T$: best rank-$k$ (eigen-)decomposition of $\mathbf{\Delta}$

Returned solution: $(\tilde{\mathbf{Q}}\mathbf{H}_k)\mathbf{\Pi}_k(\tilde{\mathbf{Q}}\mathbf{H}_k)^T$

Incremental Approximate Spectral Clustering

Guideline

- Consider the *normalized signless Laplacian* $\hat{\mathbf{L}}_t = \mathbf{I} + \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$
- Update of the best rank-k_2 approximation of $\hat{\mathbf{L}}_t$
- Initialize k-means with previous clustering

Algorithm

Require: Graphs $G_1, \ldots, G_T$, no. of clusters k_1, matrix approximation rank $k_2 \geq k_1$, eigen-decomposition recomputation step R

1: **for** $t = 1 \rightarrow T$ **do**
2: Compute the shifted Laplacian for G_t, $\hat{\mathbf{L}}_t$
3: **if** $i \,\%\, R == 1$ **then**
4: Compute rank-k_2 eigen-decomposition of $\hat{\mathbf{L}}_t$, $\mathbf{Q}_{k_2}^{(t)}\mathbf{\Omega}_{k_2}^{(t)}\mathbf{Q}_{k_2}^{(t)^T}$
5: **else**
6: Update rank-k_2 eigen-decomposition of $\hat{\mathbf{L}}_t$, $\mathbf{Q}_{k_2}^{(t)}\mathbf{\Omega}_{k_2}^{(t)}\mathbf{Q}_{k_2}^{(t)^T}$
7: **end if**
8: Let $\mathbf{Q}_{k_1}^{(t)}$ be the matrix of the first k_1 columns of $\mathbf{Q}_{k_2}^{(t)}$
9: Normalize the rows of $\mathbf{Q}_{k_1}^{(t)}$
10: Use k-means on rows of $\mathbf{Q}_{k_1}^{(t)}$, using $\mathbf{c}_{t-1}$ to find initial centroids (if not at first iteration), and store indicators $\mathbf{c}_t \in \{1, \ldots, k_1\}^{n_t}$
11: **end for**
12: **return** $\mathbf{c}_1 \in \{1, \ldots, k\}^{n_1}, \ldots, \mathbf{c}_T \in \{1, \ldots, k\}^{n_T}$

Experimental setting

Comparison to

- Exact: standard spectral clustering at each iteration
- Ning et al.: another approach updating the eigen-decomposition [3]

Two real dataset

- Sexual contacts graph [1] (2,387 nodes)
- Purchases history of an e-commerce website [4] (5,000 nodes)

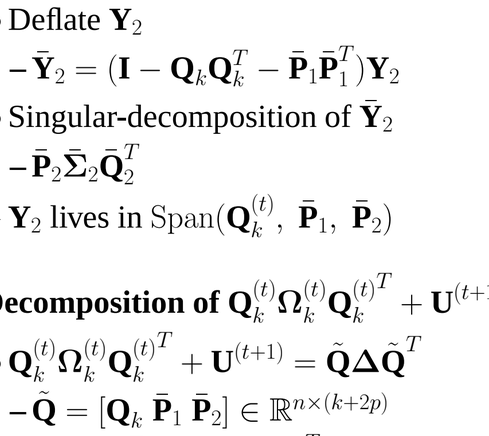
Criteria

- k-way normalized cut

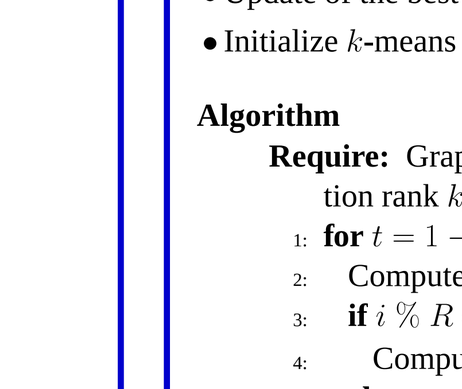
$$\frac{1}{k}\sum_{\ell=1}^{k}\frac{\sum_{ij}\mathbf{W}_{ij}\delta(\mathbf{c}_i,\ell)(1-\delta(\mathbf{c}_j,\ell))}{\sum_{ij}\mathbf{W}_{ij}\delta(\mathbf{c}_i,\ell)}$$

- The lower the better

Results



Sexual contacts



E-commerce

Remark

- `IASC` $\approx$ `Exact`
- `IASC` uses less than 10% of the eigen-vectors

Conclusion

- Fast update of the eigen-decomposition
- 10% of eigen-vectors is enough to summarize he Laplacian matrix

References

[1] B. Auvert, H. de Arazoza, S. Clémençon, J. Perez, and R. Lounes. The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low HIV prevalence. *BMC Infectious Diseases*, 7(30), November 2007.

[2] F.R.K. Chung. *Spectral graph theory*. Amer. Math. Soc., 1997.

[3] H. Ning, W. Xu, Y. Chi, Y. Gong, and T.S. Huang. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113–127, 2010.

[4] E. Richard, N. Baskiotis, T. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. In *NIPS'10*, pages 1966–1974, 2010.